Analyzing the Semantic Relatedness of Paper Abstracts

Ionut Cristian Paraschiv, Mihai Dascalu, Stefan Trausan-Matu, Philippe Dessus

Analyzing the Semantic Relatedness of Paper Abstracts

An Application to the Educational Research Field

Ionut Cristian Paraschiv, Mihai Dascalu, Stefan Trausan-Matu
Computer Science Department
University Politehnica of Bucharest
Bucharest, Romania
ionut.paraschiv@cti.pub.ro, mihai.dascalu@cs.pub.ro, stefan.trausan@cs.pub.ro

Philippe Dessus
LSE, Univ. Grenoble Alpes
France
philippe.dessus@upmf-grenoble.fr

***Abstract*—Each domain, along with its knowledge base, changes over time and every timeframe is centered on specific topics that emerge from different ongoing research projects. As searching for relevant resources is a time-consuming process, the automatic extraction of the most important and relevant articles from a domain becomes essential in supporting researchers in their day-to-day activities. The proposed analysis extends other previous researches focused on extracting co-citations between the papers, with the purpose of comparing their overall importance within the domain from a semantic perspective. Our method focuses on the semantic analysis of paper abstracts by using Natural Language Processing (NLP) techniques such as Latent Semantic Analysis, Latent Dirichlet Allocation or specific ontology distances, i.e., WordNet. Moreover, the defined mechanisms are enforced on two different subdomains from the corpora generated around the keywords "e-learning" and "computer". Graph visual representations are used to highlight the keywords of each subdomain, links among concepts and between articles, as well as specific document similarity views, or scores reflecting the keyword-abstract overlaps. In the end, conclusions and future improvements are presented, emphasizing nevertheless the key elements of our research support framework.**

Keywords—scientometrics; discourse analysis; semantic similarity; extraction of domain key concepts

I. INTRODUCTION

The problem and necessity of annotating documents from a domain grows as more and more papers and researches appear due to the fast era we live in. Having a good visual representation of the papers about a specific subject can become a major advantage for every researcher that needs to deepen his knowledge of a given topic. Moreover, a complex, interactive and intuitive visual representation of the keywords from the domain or its most relevant papers can also be used to stimulate one's imagination and even to increase the rating of new important and significant work. At the same time, the visualization can help to better understand some specific topics from a domain by providing a more comprehensive overview of the existing information.

The analyzed data in the current work is arbitrarily taken from the citation index Web of Science, from the category Education and Educational Research [1, 2], taken between the years 2000-2004, and the article abstracts are analyzed semantically as to compute the most important papers from a specific domain. It is important to note that all the domain views are applied on two subdomains from the initial data set (one containing the "*e-learning*" keyword, while the second is centered on the keyword "*computer*").

This study initially focuses on an existing paper comparison method, the co-citation analysis [3], and continues with the brief describing of the techniques used as background for current work. Later on, our experiment is described in detail along with examples of the user interface. In the end, conclusions and future work are presented.

II. CO-CITATION ANALYSIS

Co-citation analysis is a technique used frequently in the article analysis tools, usually by finding patterns of citations and by building indexing tools. An existing research [3] uses co-citation analysis to build a graph of articles in which two papers *A* and *B* are connected if they cite at least a common reference, and, in this case, the final graph contains a directed link from *A* to *B* weighted by the number of co-citations between the papers. In the created graph, different algorithms can be applied in order to find patterns in the structure (i.e., which authors cite other authors) or to find the central papers. This method is very fast to process as usually the papers from the domain are already indexed and the links between the articles are manually created by the authors. However, citation strategies are influenced by researchers' practices [4] and motivations [5]. The major disadvantage of this method is that the citations are made by the researchers and they are not necessarily accurate or sensitive (e.g., a paper cites another one more than once [5]). Moreover, a citation usually refers to, and links, only certain parts of papers not indicated by the authors, thus loosing semantic relevant meaning within the final graph.

III. SEMANTIC RELATEDNESS BETWEEN PAPER ABSTRACTS

The current research builds a graph of the papers from an initial document set by analyzing their abstracts and the semantic distances between them, using the hypothesis that the abstracts usually contain the central concepts of each paper [6]. Unlike the co-citation analysis, this model has a significant computational cost, as it needs to index all the abstracts and compute the similarity between them in order to build the *paper semantic similarity graph*.

The purpose of the research is to create different views of documents from the initial dataset as to support researchers in their activities and to provide an overview of the most important publications and concepts from a specific domain. As computing semantic distances between paper abstracts is the basis of our research, the text processing pipeline will be described later on, along with NLP methods such as Latent Semantic Analysis (LSA) [7] and Latent Dirichlet Allocation (LDA) [8] integrated within the ReaderBench system [9, 10].

Latent Semantic Analysis [7] is a model that computes the distance between two documents in a semantic vector space by extracting the most important concepts, by creating a sparse term-document matrix and by ignoring semantically void words such as stop words. A Singular Value Decomposition (SVD) is applied on the term-document matrix in order to reduce the dimensionality of the vector space. In the end, cosine similarity is applied between the associated vectors for computing the distances between documents and words.

On the other hand, Latent Dirichlet Allocation [8] is a probabilistic model also applied on large training corpora. Words are grouped based on co-occurrence patterns into topics and each document is represented as a topic distribution. The topics are classified into multiple categories and words are assigned to them, for example using their semantic meaning (e.g., "*egg*" and "*bread*" are from the "*food*" category) and have a corresponding probability to pertain to a given topic (if "*dog*" has a high probability within a topic, words like "*puppy*" or "*bone*" will also have a high probability). However, LDA can be easily applied to our problem of annotating abstracts and, combined with LSA, it improves the adequacy of the overall similarity score.

Moreover, computing semantic distances between words is of particular interest. Starting from an ontology as a conceptualization of a given domain, different metrics can be used to compute semantic distances between words [11]: number of links between nodes or types of relationships. Our system uses WordNet [12], a very large and frequently used ontology in English, containing more than 150.000 concepts. In addition, it is important to note that these measures (based on LSA, on LDA and on semantic distances in ontologies) can be combined into an aggregated cohesion score [13] as to obtain more accurate results.

Before applying the previous semantic measures, the input abstracts must go through a standard Natural Language Processing pipeline [14] in order to increase the accuracy of final similarity score. Initially, the documents must suffer some spellchecking and stop words removal, in order to keep only the most relevant semantic information. After the text is split and tokenized into vectors of words, all the components are reduced to their morphological unit using stemming (i.e., Snowball stemmer [15]) and their dictionary base form by using lemmatization. Another important task is to find named entities that can provide important semantic information for the input text that, combined with co-reference resolution [16], can increase the overall cohesion of the abstract.

Overall, our model combines all the previous methods in order to compute an aggregated cohesion score [10] and uses the most important words from the abstracts. For increased performance, only the first *N* words ordered by their relevance are used to compare the abstracts, where *N*, experimentally set at 20, is a tradeoff between accuracy and speed.

Nevertheless, the semantic analysis method encounters problems when a paper co-quotes an important document from the knowledge base, as that paper can automatically become a central article within our paper semantic similarity graph. However, potential solutions are provided in the conclusions section, along with other possible extensions of our current work.

IV. EXPERIMENTS

The purpose of the system is to create a tool that can facilitate the process of finding relevant papers and of conceptualizing a domain for a researcher. We used two corpora for the educational science domain from Jensen & Grauwin's database [1]. The first corpus that contains the keyword "*computer*" has a total number of 749 abstracts, while the one centered on "*e-learning*" has a total number of 598 files (see Table 1). Specific views are generated and used to demonstrate the sustainability of the model's results.

TABLE I. DATASET DESCRIPTION

	Education and Educational Research	
	computer	***e-learning***
No. of abstracts	749	598
Main keywords	student, computer, learning	learning, student, system
No. of central/ important articles	5	10

ReaderBench [9, 10] was modified and extended in order to perform the current analysis. *ReaderBench* is a research platform that uses a cohesion graph to represent the underlying structure of documents. It also has a dialogism module used for viewing and measuring the inter-animation of voices [17] as to determine the impact of participants' utterances within a conversation [18]. Therefore, *ReaderBench* represented a good starting point for the paper abstract model as it already had implemented many of the features described in the previous chapter. In a nutshell, by using the NLP pipeline from *ReaderBench* along with the term-document matrix representation for the abstracts, a similarity score between abstracts can be computed. The score is used along with the Gephi library [19] for representing the documents and for extracting centrality scores to reflect each paper's importance from specific Social Network Analysis measures [20]. These

are the methods used in the experiment to represent the documents and the most important concepts from the corpora.

All the forthcoming representations will be explained for both input subsets, and the most important articles along with some relevant texts from them will be extracted in order to highlight the obtained results.

A. Document Similarity View

The initial representation creates a graph with the documents connected between them if their corresponding similarity scores are higher than a selectable threshold. The size of the nodes is proportional to their centrality score.

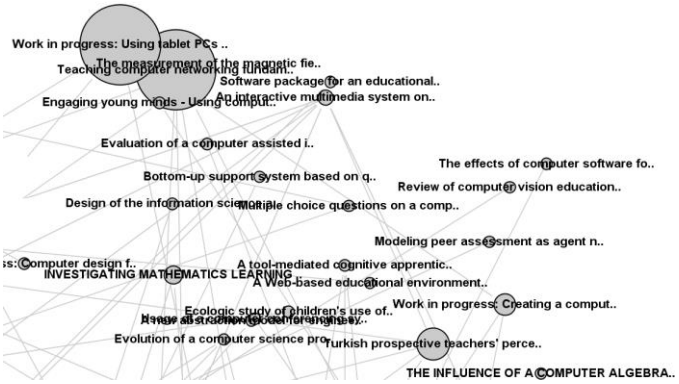

Fig. 1. Document similarity view for the "*computer*" subdomain with an imposed 75% similarity threshold

Fig. 1 represents a subsection from the graph with all the papers that contain the keyword "*computer*". As there are 749 articles and all of them contain this keyword, it is clear that there is a high semantically related even for a 75% threshold level. In this case, the central document is "*Teaching computer hardware and organisation using PIC-based projects*" („We have prepared a series of interesting projects that give students a hands-on introduction to computer hardware and organisation. Our projects, designed around the PIC16F84, a powerful 8-bit microcontroller chip that sells for less than $10, are suitable for classroom use in introductory-level courses about computer hardware."), and the most similar papers to this one are "*Using open source software in computer science courses*" and "*Integrating formal methods tools into undergraduate computer science curriculum*", titles that clearly indicate a resemblance with the concept "*computer*".

The main purpose for enabling manually adjustable thresholds with increments of 5% consists of having the posibility to finely tune the number and the density of displayed articles. Moreover, we opted to make changes between the generated networks (Figs. 1-3) in order to maximize the impact of the views and to balance the number of displayed articles in an equitable manner.

Fig. 2. Document similarity view for the "*computer*" subdomain with an imposed 90% similarity threshold

If we increase the threshold of similarity up to a minimum of 90%, a subsection from the graphs is depicted in Fig. 2 and the central paper becomes "*Complexity of integrating computer technologies into education in Turkey*" ("Integrating Information and Communication Technologies into a centralized education system such as Turkey's depends on its successful design and application, which is an expensive and complex process."), and the most similar papers to it are "*An interactive multimedia system on computer architecture, organization, and design*" and "*Evaluation of a learning repository system approach established in the schools and faculties of information technology and computer science in three large universities*".

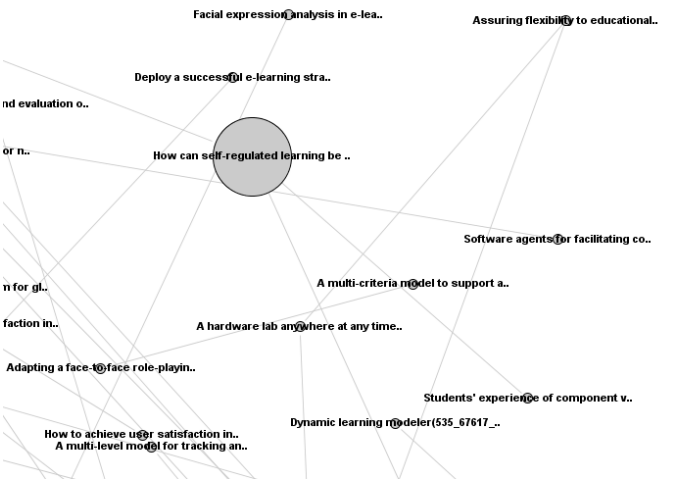

Fig. 3. Document similarity view for the "*e-learning*" subdomain with an imposed 85% similarity threshold

Fig. 3 represents a picture of the graph for the papers that contain the keyword "*e-learning*", with an imposed minimum threshold of 85%. We can check from the graph that the central paper is "*How can self-regulated learning be supported in mathematical E-learning environments?*" ("This

study compares two E-learning environments: E-learning supported with IMPROVE self-metacognitive questioning (EL+IMP), and E-learning without explicit support of self-regulation (EL).“). The most similar papers are “*An architecture of virtual environment for e-learning*“ and “*Students' experience of component versus integrated virtual learning environments*“.

These views demonstrate that the most central papers are within the semantic context induced by the selected keywords. Also, as the number of displayed articles changes with different threshold values, the most central papers dynamically change for each generated view. However, they are not necessarily the most important articles from the domain as the identified papers have their abstracts closest to the greatest number of other papers' abstracts. However, there is a problem with this representation: if an outside article having a very similarity to a central graph paper is added, it would automatically be interpreted as one of the central nodes.

B. Document Space View for a Particular Paper

A different approach to observe a research field is to find the most similar documents for an initial, user selected paper. In this situation, the Document Space View was created, as to further extend an initial paper conceptualization with similar articles. This view has a maximum depth level of 3, showing on the first level all the documents similar over a threshold with the initial one, and on the next levels using all the documents from the previous step as initial papers in order to try to find new documents. The imposed similarity threshold is selectable by the user, giving the chance to differently model the level of similarity between the papers from the graph.

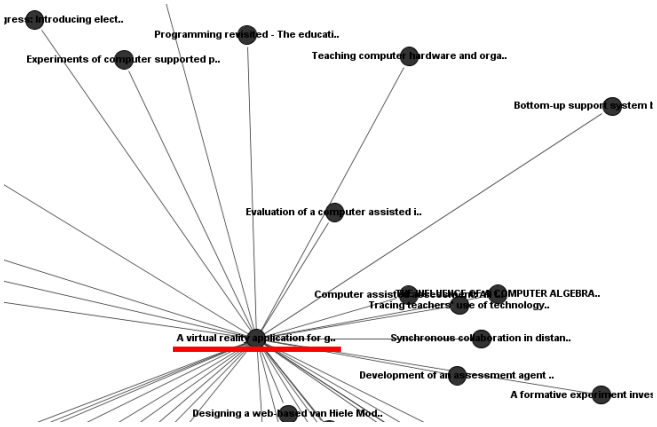

Fig. 4. Document space view for a selected paper with an imposed 70% similarity threshold and a depth level of 1

Fig. 4 shows the graph for the explorative view of paper “*A virtual reality application for geometry classes*“ (“This paper describes the use of the Virtual Reality Modelling Language (VRML) to visualise 3-D objects for middle school geometry classes in a networked environment and shows its usefulness for both teacher and students.“), selected to be the central paper (the article contains the keyword “*computer*“). For a depth level of 1 and a 70% threshold, the most similar papers to it from the graph are “*Bottom-up support system based on query pattern in computer education*“ (“In this work, we developed a query support system that is efficient for learners who are studying after lessons. This system can serve for students to enhance their abilities to deal with problems by themselves and can serve for teachers to offer important information to help improve class.“) and “*Synchronous collaboration in distance education: A case study on a computer science course*“ (“This paper describes our experience with introduction of synchronous collaborative problem solving activities in the frame of a distance learning computer science undergraduate course of the Hellenic Open University“).

On the other hand, Fig. 5 depicts the same graph for a depth level of 2 and an 80% threshold. It is important to note that the paper selected to build the graph is not the most central paper any more, as we can clearly see from the small node’s size. Even though the threshold from Fig. 5 is higher than in Fig. 4, we can see how many new nodes and links are being added to the graph. To better visualize parts from the graph, it needs to be zoomed in or the threshold needs to be increased. In Fig. 5, the most related papers to the central one are the same as in Fig. 4, and they are the same for every depth level.

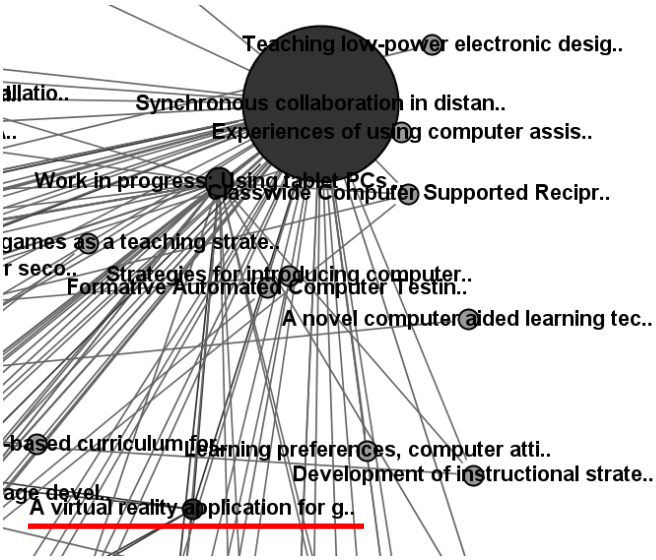

Fig. 5. Document space view for a selected paper with an imposed 80% similarity threshold and a depth level of 2

C. Concept Map View

A helpful facility to stimulate creativity is to display a graph with the most important concepts from the field. Starting from this viewpoint, the concept view contains a graph with the most relevant and important concepts from the document set, where the nodes (words) are sized according to their centrality. In other words, this view can help researchers by presenting different important words from which they can select appropriate subsets of papers that contain concepts or similar words that they want to explore.

Through the use of *ReaderBench*, all the important words from the abstract are extracted, and an average score for every word across all the documents is created. From these concepts, the first 50 are extracted and a similarity matrix between them

is created using *WordNet*. The resulting graph is generated in a similar manner to the previous methods: two words are connected if their similarity exceeds the threshold, and their betweenness centrality used to define the size of each node increases when they become more connected.

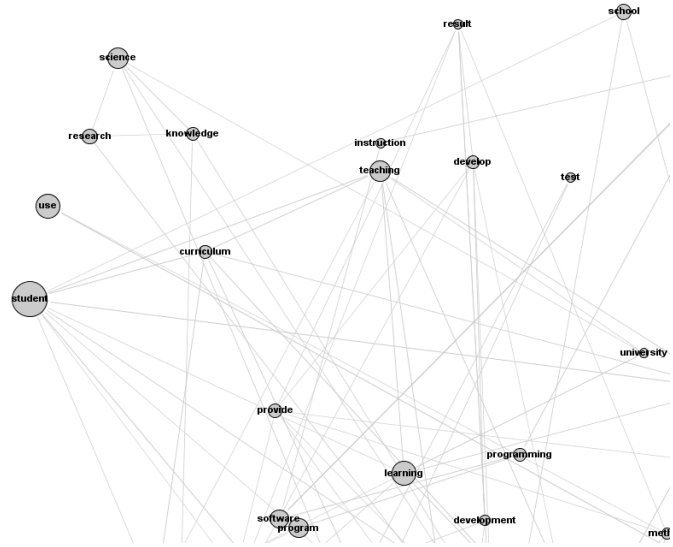

Fig. 6. Concept map of "*computer*" papers

Fig. 6 displays the most important words for the papers that contain "*computer*" as their keyword. For a threshold of 60%, the words ordered by their relevance are "*student*", "*computer*", "*learning*", "*study*", "*system*", and "*teaching*". This result proves the effectiveness of our method, as "*computer*" is one of the most important words from the document set.

Fig. 7 represents the word graph for the "*e-learning*" document subset. The most important words extracted for this subset are "*learning*", "*student*", *system*", "*learn*", "*course*", "*education*", "*study*", "*process*", "*model*", "*technology*", that are clearly in the semantic context of *e-learning*.

This approach confirms that the proposed semantic analysis can lead to an adequate representation and understanding of a domain and can generate a general perspective of the withheld knowledge.

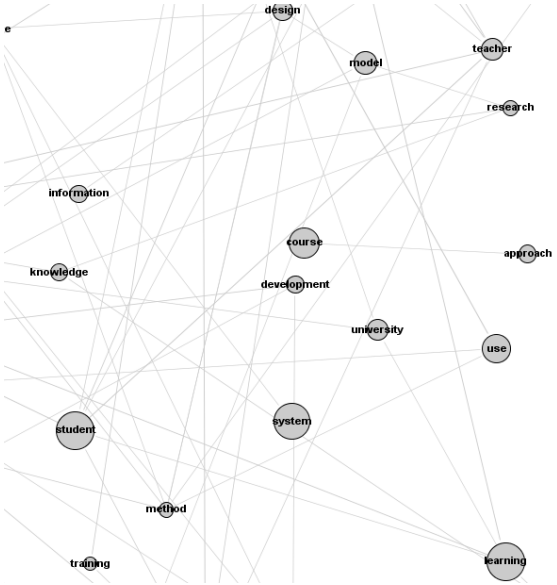

Fig. 7. Concept map of "*e-learning*" papers

D. Keyword-abstract overlap

The keyword-abstract overlap view computes a relevance score between the keywords and the abstract of each paper. The relevance is computed as a compound score between the number of occurrences of keywords in the abstract (30%) and their semantic relatedness (70%). The previous weights were empirically assigned to best reflect the overall similarity while considering both lexical and semantic relatedness.

Article: Instructional design model promoting transfer using group development method of e-learning teaching materials by learners themselves
Abstract: Truly useful classes are that the knowledge and skills learned can be applied to other learning domains in the future. This concept is called "Transfer" in cognitive science. This paper adopts the viewpoints of Seneca as leading concepts. Based on these concepts, we have developed an instructional design model promoting Transfer using a method where learners develop e-Learning teaching materials themselves as a group. e also confirm the validity of this model. The learning procedure of this model uses a double loop structure in a learning cycle. The double loop structure consists of "repetition of judgment concerning instructional intention" during "repetition of group development process throughout the course". The steps of this learning cycle are as follows. 1) Project Choice, 2) Grouping of Learners, 3) Solution Search, 4) Group Development of e-Learning Teaching Materials, 5) Rehearsals of Presentation and Judgment Concerning Instructional Intention, 6) Presentation of Solutions, 7) Suggestion of Improvements and Transfer Promotion, and 8) Making of Legacy Teaching Materials end of the course. The validity of this instructional design model in promoting Transfer is confirmed in the results of causality analysis on learner questionnaires (by structural equation modeling). In addition, we expect to reduce the e-Learning teaching materials development load on teachers because of the characteristics of the learning cycle.
Keywords: group development method, instructional design model, legacy, metacognition, structural equation modeling (sem), transfer
Syntactic overlap: *1.00*; Semantic overlap: *0.69*; Aggregated score: *0.79*

Fig. 8. Keyword-abstract overlap for a highly relevant paper abstract with correctly assigned keywords

The computed similarity measure between the abstract and the keywords reflects overall a reliable estimation of the adequacy of the manually defined keywords in contrast to the actual abstract. In Fig. 8, the article with the highest score (0.79) is very relevant, as most of its keywords are inside the content of the abstract and the two subsets are semantically related.

Article: E-learning: Do our students want it and do we care?
Abstract: Early childhood courses at the University of Western Sydney are at a watershed. Program restructuring has embraced the challenges of the changing contexts of Australian early childhood education and the dynamic multicultural, multilingual,

multi-aged communities of Western Sydney. These conditions have resulted in the reconceptualisation of the content and delivery of initial and continuing education for early childhood professionals at UWS. This paper will present research conducted by the early childhood staff team as they document and analyse the introduction of new courses using a blended learning approach.
Keywords: technologies for marginalised and disadvantaged
Syntactic overlap: *0.00*; Semantic overlap: *0.26*; Aggregated score: *0.19*

Fig. 9. Keyword-abstract overlap for a lesser relevant abstract, quite short in terms of length and with generic keywords

In contrast, the articles from the last places usually have a short and unclear abstract, with a limited number of keywords not related to the actual textual description, or not that descriptive (see Fig. 9 for *.19* score). Moreover, this score can be used to automatically propose adequate keywords for an article that uses too broad or irrelevant words within the provided list. The examples in this subchapter where taken from the "*e-learning*" subset.

V. Conclusions and Future Work

In order to argue for the feasibility of our model, full text analysis methods were applied on two subsets of documents extracted from the Web of Science citation index, Education and Educational Research category, centered on two keywords: "*computer*" and "*e-learning*". *ReaderBench* can be used to explore networks of papers graphically, by domain, based on the semantic relatedness of their abstracts rather than by co-citation.

As envisioned extensions, it can be extremely interesting to check the citation motivation of the author by adding sentiment analysis [21] or by computing inter-paragraph cohesion [22] to display the paper's semantic flow. In order to increase the reliability and quality of the final views, the document scores should be weighted using the co-citation approach. Moreover, another possible extension is to display the paper graphs using a time frame, view in which researchers can better observe the trending topics or the central articles.

All in all, the semantic analysis of paper abstracts is a good start for annotating papers with semantic metadata and for increasing the general representation and visualization of the key concepts within a given domain.

Acknowledgments

The work presented in this paper was partially funded by the Sectorial Operational Programme Human Resources Development 2007-2013 of the Ministry of European Funds through the Financial Agreement POSDRU/159/1.5/S/134398. We also thank Pablo Jensen and Sebastian Grauwin for providing the initial corpus of paper abstracts.

References

[1] S. Grauwin and P. Jensen, "Mapping scientific institutions," *Scientometrics*, 2011.

[2] S. Grauwin and P. Jensen. EducMap Project [Online]. Available: http://ife.ens-lyon.fr/ife/recherche/groupes-de-travail/educmap/educmap

[3] K. W. Boyack and R. Klavans, "Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately?," *Journal of the American Society for Information Science and Technology*, vol. 61, pp. 2389–2404, 2010.

[4] H. Sword, *Stylish Academic Writing*. Cambridge, Massachusetts & London, England: Harvard University Press, 2012.

[5] M. Song and Y. Ding, "Topic modelling: Measuring scholarly impact through the topical lens " in *Measuring scholarly impact: Methods and practice*, Y. Ding, R. Rousseau, and D. Wolfram, Eds., ed: Springer, 2014, p. 346.

[6] Y. Ding, M. Song, X. Wang, G. Zhang, Z. C., and T. Chambers, "Content-based citation analysis: The next generation of citation analysis," *Journal of the American Society for Information Science & Technology*, vol. 65, 2014.

[7] T. K. Landauer and S. T. Dumais, "A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge," *Psychological Review*, vol. 104, pp. 211–240, 1997.

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[9] M. Dascalu, *Analyzing discourse and text complexity for learning and collaborating*, *Studies in Computational Intelligence* vol. 534. Switzerland: Springer, 2014.

[10] M. Dascalu, P. Dessus, M. Bianco, S. Trausan-Matu, and A. Nardy, "Mining texts, learners productions and strategies with ReaderBench," in *Educational Data Mining: Applications and Trends*, A. Peña-Ayala, Ed., ed Switzerland: Springer, 2014, pp. 335–377.

[11] A. Budanitsky and G. Hirst, "Evaluating WordNet-based Measures of Lexical Semantic Relatedness," *Computational Linguistics*, vol. 32, pp. 13–47, 2006.

[12] G. A. Miller, "WordNet: A lexical database for English," *Communications of the ACM*, vol. 38, pp. 39–41, 1995.

[13] M. Dascalu, P. Dessus, S. Trausan-Matu, M. Bianco, and A. Nardy, "ReaderBench, an environment for analyzing text complexity and reading strategies," in *AIED 201)*, Memphis, USA, 2013, pp. 379–388.

[14] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*, 2nd ed. London: Pearson Prentice Hall, 2008.

[15] M. Porter and R. Boulton, "Snowball," ed. http://snowball.tartarus.org/, 2002.

[16] H. Lee, A. Chang, Y. Peirsman, N. Chambers, M. Surdeanu, and D. Jurafsky, "Deterministic coreference resolution based on entity-centric, precision-ranked rules," *Computational Linguistics*, vol. 39, 2013.

[17] S. Trausan-Matu, G. Stahl, and J. Sarmiento, "Polyphonic Support for Collaborative Learning," in *CRIWG 2006*, Medina del Campo, Spain, 2006, pp. 132–139.

[18] M. Dascalu, S. Trausan-Matu, and P. Dessus, "Validating the Automated Assessment of Participation and of Collaboration in Chat Conversations," in *ITS 2014*, Honolulu, USA, 2014, pp. 230–235.

[19] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: An open source software for exploring and manipulating networks," in *Int. AAAI Conf. on Weblogs and Social Media*, San Jose, CA, 2009, pp. 361–362.

[20] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*. Cambridge, UK: Cambridge University Press, 1994.

[21] D. Lupan, M. Dascalu, S. Trausan-Matu, and P. Dessus, "Analyzing emotional states induced by news articles with Latent Semantic Analysis," in *AIMSA 2012*, Varna, Bulgaria, 2012, pp. 59–68.

[22] S. T. O'Rourke and R. A. Calvo, "Analysing semantic flow in academic writing," in *AIED2009*, V. Dimitrova, R. Mizoguchi, B. du Boulay, and A. C. Graesser, Eds., ed Amsterdam, The Netherlands: IOS Press, 2009, pp. 173–180.